C4.5 ALGORİTMASI

NİTELİK1	NİTELİK2	NİTELİK3	SINIF
a	70	DOĞRU	S1
a	90	DOĞRU	S2
a	85	YANLIŞ	S2
a	95	YANLIŞ	S2
a	70	YANLIŞ	S1
b	90	DOĞRU	S1
b	75	YANLIŞ	S1
b	65	DOĞRU	S1
b	75	YANLIŞ	S1
c	80	DOĞRU	S2
c	70	DOĞRU	S2
c	80	YANLIŞ	S1
c	70	YANLIŞ	S1
c	96	YANLIŞ	S1

Adım 1 : NİTELİK2'nin değerleri küçükten büyüğe ve her birinden sadece bir kez olacak şekilde sıralanır ve orta noktalar bulunur.

Nitelik2 = { 65, 70, 75, 80, 85, 90,95,96 } için orta noktalar (80 , 85) 'tir.

En uygun " t eşik değeri " hesaplanır.

t = Orta noktaların toplamı / 2 O hâlde t = 80+85 / 2 = 82.5 yani **83** diye alınır.

Adım 2 : NİTELİK2'nin değerleri artık **83'ten büyükler** ve **83'ten küçükler** şeklinde iki gruba ayrılır.

NİTELİK1	NİTELİK2	NİTELİK3	SINIF
a	83'ten küçük	DOĞRU	S1
a	83'ten büyük	DOĞRU	S2
a	83'ten büyük	YANLIŞ	S2
a	83'ten büyük	YANLIŞ	S2
a	83'ten küçük	YANLIŞ	S1
b	83'ten büyük	DOĞRU	S1
b	83'ten küçük	YANLIŞ	S1
b	83'ten küçük	DOĞRU	S1
b	83'ten küçük	YANLIŞ	S1
c	83'ten küçük	DOĞRU	S2
c	83'ten küçük	DOĞRU	S2
c	83'ten küçük	YANLIŞ	S1
c	83'ten küçük	YANLIŞ	S1
c	83'ten büyük	YANLIŞ	S1

Önce hedef sınıf olan SINIF niteliğinin entropisi hesaplanır. S1'den 9 ve S2'den 5 tane mevcut.

H(SINIF) = - (9/14*$\log_2$9/14 + 5/14*$\log_2$5/14) = 0.94

Adım 3 : NİTELİK2'nin nitelik değerleri için SINIF niteliğine göre entropiler hesaplanır.

83'ten küçük olanların sayısı : 9 tane.

7 tanesi S1, 2 tanesi S2 sınıfında.

H(NİTELİK2 $_{<83}$) = - (7/9*$\log_2$7/9 + 2/9*$\log_2$2/9) =0.76

83'ten büyük olanların sayısı : 5 tane.

2 tanesi S1, 3 tanesi S2 sınıfında.

H(NİTELİK2 >83) = - (2/5*$\log_2$2/5 + 3/5*$\log_2$3/5) = 0.97

H(NİTELİK2, SINIF) = 9/14 * H(NİTELİK $_{<83}$) + 5/14 * H(NİTELİK2 $_{<83}$) = 0.84

Adım 4 : Kazanç ölçütü hesaplanır.

Kazanç (NİTELİK2, SINIF) = H(SINIF) – H(NİTELİK2, SINIF) = 0.1

Bilinmeyen / Kayıp veri hesaplama :

NİTELİK1	NİTELİK2	NİTELİK3	SINIF
a	70	DOĞRU	S1
a	90	DOĞRU	S2
a	85	YANLIŞ	S2
a	95	YANLIŞ	S2
a	70	YANLIŞ	S1
?	90	DOĞRU	S1
b	75	YANLIŞ	S1
b	65	DOĞRU	S1
b	75	YANLIŞ	S1
c	80	DOĞRU	S2
c	70	DOĞRU	S2
c	80	YANLIŞ	S1
c	70	YANLIŞ	S1
c	96	YANLIŞ	S1

F = bilinen verilerin sayısı / TOPLAM için yeni kazanç ölçütü **Kazanç(X, T) = F * [H(T) – H(X, T)]**

Adım 1 : Bilinmeyen verinin olduğu satır veri setinden komple çıkarılır.

F = 13 / 14 = 0.93

NİTELİK1	NİTELİK2	NİTELİK3	SINIF
a	70	DOĞRU	S1
a	90	DOĞRU	S2
a	85	YANLIŞ	S2
a	95	YANLIŞ	S2
a	70	YANLIŞ	S1
b	75	YANLIŞ	S1
b	65	DOĞRU	S1
b	75	YANLIŞ	S1
c	80	DOĞRU	S2
c	70	DOĞRU	S2
c	80	YANLIŞ	S1
c	70	YANLIŞ	S1
c	96	YANLIŞ	S1

Adım 2 : Yeni entropi hesaplanır. S1'den 8 tane, S2'den 5 tane mevcut.

H(SINIF) = - ($5/13*\log_2 5/13$ + $8/13*\log_2 8/13$) = 0.96

Adım 3 : NİTELİK1 için SINIF niteliğine göre entropi hesaplanır.

NİTELİK1'de **5** tane a, **3** tane b, **5** tane de c değeri var.

a'lardan 2 tanesi S1, 3 tanesi de S2 sınıfında. H(NİTELİK1 a) = - ($2/5*\log_2 2/5$ + $3/5*\log_2 3/5$) = 0.97

b'lerden 3 tanesi S1, 0 tanesi de S2 sınıfında. H(NİTELİK1 b) = - ($3/3*\log_2 3/3$ + $0/3*\log_2 0/3$) = 0

c'lerden 3 tanesi S1, 2 tanesi de S2 sınıfında. H (NİTELİK1 c) = - ($3/5*\log_2 2/5$ + $2/5*\log_2 2/5$) = 0.97

Adım 4 : H(NİTELİK1, SINIF) değeri bulunduktan sonra yeni kazanç hesaplanır.

H(NİTELİK1, SINIF) = 5/13 * **H(NİTELİK1 a)** + 3/13 * **(NİTELİK1 b)** + 5/13 * **H(NİTELİK1 c)** = 0.14

Kazanç(NİTELİK1, SINIF) = F * [H(SINIF) – H(NİTELİK1, SINIF)] = 0.93 * (0.96 – 0.14) = **0.76**

Hakan Cem Gerçek

gmail: hakancg95gmail.com

instagram : hkn.cem

Twitter: eightjune95